Entropy and Information Causality in General Probabilistic Theories

Howard Barnum,[1,*] Jonathan Barrett,[2,†] Lisa Orloff Clark,[3,‡] Matthew Leifer,[1,4,§]
Robert Spekkens,[1,¶] Nicholas Stepanik,[3,**] Alex Wilce,[3,††] and Robin Wilke[5,‡‡]

[1]*Perimeter Institute for Theoretical Physics, 31 Caroline St N, Waterloo, Ontario, N2L 2Y5 Canada*

[2]*H. H. Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol, BS8 1TL United Kingdom*

[3]*Department of Mathematical Sciences, Susquehanna University, Selinsgrove, PA, 17870 USA*

[4]*Institute for Quantum Computing, University of Waterloo, 200 University Ave. W, Waterloo, Ontario, N2L 3G1 Canada*

[5]*Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05405 USA*

(Dated: September 28, 2009)

We investigate the concept of entropy in probabilistic theories more general than quantum mechanics, with particular reference to the notion of information causality recently proposed by Pawlowski *et al.* (arXiv:0905.2992). We consider two entropic quantities, which we term *measurement* and *mixing* entropy. In the context of classical and quantum theory, these coincide, being given by the Shannon and von Neumann entropies respectively; in general, however, they are very different. In particular, while measurement entropy is easily seen to be concave, mixing entropy need not be. In fact, as we show, mixing entropy is not concave whenever the state space is a non-simplicial polytope. Thus, the condition that measurement and mixing entropies coincide is a strong constraint on possible theories. We call theories with this property *monoentropic*.

Measurement entropy is subadditive, but not in general strongly subadditive. Equivalently, if we define the mutual information between two systems A and B by the usual formula $I(A:B) = H(A) + H(B) - H(AB)$ where H denotes the measurement entropy and AB is a non-signaling composite of A and B, then it can happen that $I(A:BC) < I(A:B)$. This is relevant to information causality in the sense of Pawlowski *et al.*: we show that any monoentropic non-signaling theory in which measurement entropy is strongly subadditive, and also satisfies a version of the Holevo bound, is informationally causal, and on the other hand we observe that Popescu-Rohrlich boxes, which violate information causality, also violate strong subadditivity. We also explore the interplay between measurement and mixing entropy and various natural conditions on theories that arise in quantum axiomatics.

I. INTRODUCTION

One can view quantum mechanics as an extension of the classical probability calculus, allowing for random variables that are not simultaneously measurable. In order to gain a clearer understanding of quantum theory from this perspective, it is useful to contrast it with various (factitious) alternatives that are neither classical nor quantum. The best known example of such a "foil" probabilistic theory is probably the theory of "non-local boxes" [1, 28, 29]; but in fact, there is a standard mathematical framework for such theories, going back to the work of Mackey in the 1950s [26]. Working in this framework, one can show that many phenomena commonly regarded as characteristically quantum – no-cloning and no-broadcasting theorems [2, 3], the trade-off between state disturbance and measurement [1], and the existence and basic properties of entangled states [1, 2, 23, 24] – are actually quite generic features of all non-classical probabilistic theories satisfying a basic "non-signaling" constraint. Other quantum phenomena, such as the possibility of teleportation [4] or remote steering of ensembles [6], are more special (and in some sense, more *classical*), but can still be seen to arise outside the boundaries of quantum theory.

One might hope to find some reasonably short list of probabilistic or information-theoretic phenomena that more cleanly separate quantum theory from other possible non-signaling theories. In a recent paper [27], Pawlowski *et al.* take a step in this direction by

*Electronic address: hbarnum@perimeterinstitute.ca
†Electronic address: j.barrett@bristol.ac.uk
‡Electronic address: clarklisa@susq.edu
§Electronic address: matt@mattleifer.info
¶Electronic address: rspekkens@perimeterinstitute.ca
**Electronic address: stepanik@susqu.edu
††Electronic address: wilce@susqu.edu
‡‡Electronic address: rwilke@uvm.edu

showing that any non-signaling correlation violating the Tsirel'son bound also violates a qualitative information-theoretic principle they call *information causality* (IC). In essence, this prohibits a form of "multiplexing" in which one party (Bob) gains the ability to access a total of more than m bits of information held by another party (Alice), on the basis of an m-bit message from Alice, plus some shared non-signaling bipartite state. It is also established in [27] that quantum mechanics – and hence, also classical probability theory – satisfies this IC constraint.

In establishing that quantum mechanics satisfies IC, Pawlowski *et al.* make use only of standard formal properties of the von Neumann entropy of joint quantum states. This raises the obvious question of where their proof breaks down in other contexts (e.g., a PR box) in which IC fails. In order to address this question, we develop some of the basic machinery of entropy, conditional entropy and mutual information in a very general probabilistic setting — an independently interesting problem, which seems not to have received much previous attention (an exception being the paper [20] of Hein).

We begin by identifying two notions of entropy, which we call *measurement* and *mixing* entropy, and which we denote respectively by $H(A)$ and $S(A)$, where A is a general probabilistic model. Briefly: the measurement entropy of a system is the minimum Shannon entropy of any possible measurement thereon, while the mixing entropy is the infimum of the Shannon entropies of the various ways of preparing the system's state as a mixture of pure states. These coincide classically and in quantum theory, but are generally quite different animals. For example, measurement entropy is always subadditive, and is concave; mixing entropy is generally neither. In fact, in Appendix A, we show that there are *always* violations of concavity of the mixing entropy for any system with a state space that is a non-simplicial polytope. Thus, the condition that mixing and measurement entropies *do* coincide, as in quantum mechanics, is a powerful constraint on the structure of a probabilistic theory. We call theories with this feature *monoentropic*.

Next, we develop an account of joint measurement entropy, conditional entropy, and mutual information for composite systems, and apply this apparatus to the notion of information causality given in [27]. Somewhat surprisingly, it seems that the main issue is not so much one of the strength of non-local correlations, but rather, the failure, of two other, very basic principles. One is the strong subadditivity, or, equivalently, the condition that the mutual information, defined by $I(A:B) = H(A) + H(B) - H(AB)$, satisfy

$$I(A:BC) \geq I(A:C).$$

This holds both classically and in quantum theory, but is violated in very simple non-classical models – even models in which A and B are *classical*, so that no issue of non-locality can arise. Another basic principle, equivalent to the Holevo bound, is that $I(E:B) \leq I(A:B)$ where E is any particular measurement on system A.

Both strong subadditivity and the Holevo bound can be viewed as special cases of an even more basic principle, usually called the *data processing inequality*. This asserts that, for any systems A, B and B', and any reasonable process $\mathcal{E} : B \to B'$, we have $I(A : \mathcal{E}(B)) \leq I(A : B)$ (where $\mathcal{E}(B) := B'$ is the output system of the process). This is intuitively appealing as a basic physical postulate.

Finally, we apply the apparatus just described to the notion of information causality. We consider in detail the basic example, due to van Dam [34] of an IC-violating composite system, and find that it exhibits a violation of strong subadditivity. We also establish that, within a very broad class of finite-dimensional *monoentropic* theories, strong subadditivity together with the Holevo bound entail information causality. It remains an open question whether all three of these conditions are necessary for this conclusion.

The remainder of this paper is organized as follows. In Section II, we review in some detail the framework of generalized probability theory, largely following [2]. In section III, we define, and establish some elementary properties of, measurement and mixing entropy for states of an arbitrary probabilistic model. Section IV discusses composite systems in our framework, and collects some observations about the behavior of joint measurement entropy, and the notion of mutual information based on this. Using this apparatus, we establish in Section V that any monoentropic probabilistic theory in which measurement entropy is strongly subadditive and satisfies the Holevo bound, is informationally causal in the sense of [27]. We also point out that violations of strong subadditivity are possible in theories having no entanglement. Section VI collects some final remarks and open questions. Appendix A contains the proof that mixing entropy is not concave on state spaces that are non-simplicial polytopes. Appendix B establishes some further properties of monoentropic theories, relevant to axiomatic characterizations of quantum theories, and also shows that monoentropicity follows from two other properties, steering and pure conditioning, the physical content of which may be more transparent. Finally in Appendix C we discuss how the framework of this paper relates to the "convex sets" framework, and consider analogous definitions of measurement entropy in that context.

II. GENERAL PROBABILISTIC MODELS

As we mentioned above, there is a more or less standard mathematical framework for discussing general

probabilistic models, going back at least to the work of Mackey in the 1950s, and further developed (or, in some cases, rediscovered) in succeeding decades by various authors [1, 13, 15, 16, 18, 25]. In what follows, we work in the idiom of [8], which we briefly recall.

We characterize a probabilistic model, or, more briefly, a *system*, by a pair $A = (\mathfrak{A}, \Omega)$ where $\mathfrak{A}$ is a collection – possibly infinite – of discrete classical experiments or measurements, and Ω is a set of *states*. We make the following assumptions:

(i) Every experiment in $\mathfrak{A}$ is defined by its set of possible outcomes, so that we may represent $\mathfrak{A}$, mathematically, as a collection of sets $E, F, ...$ In the language of [16, 35], this is a *test space*; accordingly, we refer to the various sets $E, F, ... \in \mathfrak{A}$ as *tests*.

(ii) Every state is entirely determined by the probabilities it assigns to the outcomes of the various measurements in $\mathfrak{A}$. Thus, letting $X := \bigcup \mathfrak{A}$ denote the total *outcome space* of $\mathfrak{A}$, Ω consists of functions $\alpha : X \to [0, 1]$, with $\sum_{x \in E} \alpha(x) = 1$ for every set $E \in \mathfrak{A}$.

(iii) The state space Ω is a convex subset of $[0, 1]^X$ (the functions from X to $[0, 1]$). Hence any statistical mixture of states is a state.

For a given test space $\mathfrak{A}$, one can define the space of *all* states on $\mathfrak{A}$. This is called the *maximal* state space and is denoted by $\Omega(\mathfrak{A})$. It is clearly convex. The *physical* state space Ω is necessarily either equal to or a subset of the maximal state space.

This framework, though very simple, is broad enough to accommodate both measure-theoretic classical probability theory and non-commutative probability theory based on von Neumann algebras.[40] In this paper, we shall be interested exclusively in discrete, finite-dimensional systems. Accordingly, from this point forward, we make the standing assumptions that (i) $\mathfrak{A}$ is *locally finite*, meaning that all tests $E \in \mathfrak{A}$ are finite sets [41], and (ii) Ω is finite dimensional and closed.

As is easily checked, local finiteness guarantees that the maximal state space $\Omega(\mathfrak{A})$ is compact; thus, the closedness of the physical statespace Ω insures that it, too, is compact.[42] It follows that every state can be represented as a finite convex combination, or mixture, of *pure states*, that is, extreme points of Ω.

We now consider several examples. For us, a *classical system* corresponds to a pair $(\{E\}, \Delta(E))$ where the test space $\{E\}$ consists of a single measurement and where $\Delta(E)$ denotes the entire simplex of probability weights on E. In other words, there is just one test and any probability distribution over the outcomes is a possible state. A quantum system corresponds to $(\mathfrak{F}(\mathbf{H}), \Omega(\mathbf{H}))$, where $\mathfrak{F}(\mathbf{H})$ is the set of (unordered) orthonormal bases on a complex Hilbert space $\mathbf{H}$ and $\Omega(\mathbf{H})$ is the set of density operators.[43]

A simple example that is neither classical nor quantum, and to which we shall refer often, is the "two-bit" test space $\mathfrak{A}_2 = \{\{a, a'\}, \{b, b'\}\}$, consisting of a pair of two-outcome tests, depicted in Fig. 1. The full state space $\Omega(\mathfrak{A}_2)$ is isomorphic to the unit square $[0, 1]^2$ under the map $\alpha \mapsto (\alpha(a), \alpha(b))$ and is depicted if Fig. 2. Accordingly, we shall call a system of this form a *square bit* or *squit*. A PR box is a particular entangled state of two squits, as discussed below in Section V A.

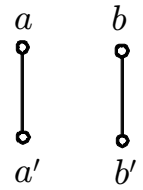

FIG. 1: The "two-bit" test space $\mathfrak{A}_2 = \{\{a, a'\}, \{b, b'\}\}$. It is depicted using a *Greechie diagram*, wherein vertices denote outcomes and every smooth line through a set of vertices represents a test.

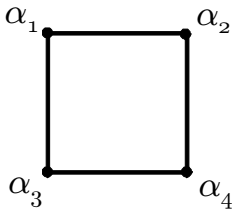

FIG. 2: The squit state space $\Omega(\mathfrak{A}_2)$. The pure state α_1 yields the outcome a in test $\{a, a'\}$ and the outcome b in $\{b, b'\}$, that is, $\alpha_1(a) = 1$, $\alpha_1(a') = 0$, and $\alpha_1(b) = 1$, $\alpha_1(b') = 0$. Similarly, α_2, α_3 and α_4 yield the pairs of outcomes a, b', a', b and a', b' respectively.

III. MEASUREMENT AND MIXING ENTROPIES

Let $\mathbf{H}$ be a finite-dimensional Hilbert space, representing a quantum system. The von Neumann entropy of a state ρ on this system is defined as $-\mathrm{Tr}(\rho \log \rho)$, where here and elsewhere, logarithms have base 2. Equivalently, it is the Shannon entropy of the coefficients λ_i in the spectral decomposition $\rho = \sum_i \lambda_i P_i$ (where the P_i are ρ's rank-one eigenprojections). In effect, the spectral decomposition privileges a particular convex decomposition of the state, and (up to phases) a privileged test in $\mathfrak{F}(\mathbf{H})$.

In our much more general setting, where we have nothing like a spectral theorem, how might we define the entropy of a state? The following definitions suggest themselves.

Definition 1 *Let α be a state on $\mathfrak{A}$. For each test $E \in \mathfrak{A}$, define the* local measurement entropy *of α at E, $H_E(\alpha)$, to be the classical (Shannon) entropy of $\alpha|_E$, i.e.,*

$$H_E(\alpha) := -\sum_{x \in E} \alpha(x) \log(\alpha(x)).$$

The measurement entropy *of α, $H(\alpha)$, is the infimum of $H_E(\alpha)$ as E ranges over $\mathfrak{A}$, i.e.,*

$$H(\alpha) := \inf_{E \in \mathfrak{A}} H_E(\alpha).$$

Note that the measurement entropy of a state of $A = (\mathfrak{A}, \Omega)$ depends entirely on the structure of $\mathfrak{A}$, and is independent of the choice of state space Ω. It will often be convenient to write $H(\alpha)$ as $H(A)$, where context makes clear which state is being considered.

For the remainder of this paper we make, and shall make free use of, the assumption that the measurement entropy of a state is actually achieved on some test, i.e., that $H(\alpha) = H_E(\alpha)$ for some $E \in \mathfrak{A}$. This is the case in quantum theory, and can be shown to hold much more generally, given some rather weak analytic requirements on an abstract model $(\mathfrak{A}, \Omega)$ – for details, see Appendix B. It follows that $H(\alpha) = 0$ if and only if there is a test such that α assigns probability 1 to one of its outcomes.

Definition 2 *Let α be a state on $\mathfrak{A}$. The* mixing *(or* preparation*) entropy for α, denoted $S(\alpha)$, is the infimum of the classical (Shannon) entropy $H(p_1, ..., p_n)$ over all finite convex decompositions $\alpha = \sum_i p_i \alpha_i$ with α_i pure.*

Again, we write $S(A)$ for $S(\alpha)$ where α belongs to the state space Ω of a system $A = (\mathfrak{A}, \Omega)$. In contrast to measurement entropy, the mixing entropy of a state depends only on the geometry of the state space Ω, and is independent of the choice of test space $\mathfrak{A}$. The mixing entropy of a pure state is 0.

Trivially, in classical probability theory, measurement and mixing entropies coincide, both being simply the Shannon entropy. Much less trivially, measurement and mixing entropies also coincide in quantum theory, where they equal the von Neumann entropy.[44] As the following example shows, however, measurement and mixing entropies can be quite different.

Example 1 *(The firefly model [45]) Let $\mathfrak{A} = \{\{a, x, b\}, \{b, y, c\}, \{c, z, a\}\}$. This test space is depicted in Fig. 3. One can check that $\Omega(\mathfrak{A})$ has five pure states, one of which is given by $\alpha(a) = \alpha(b) = \alpha(c) = 1/2, \alpha(x) = \alpha(y) = \alpha(z) = 0$: since this is pure, $S(\alpha) = 0$, yet $H(\alpha) = 1$. On the other hand, consider the pure states β and γ determined by $\beta(b) = \beta(z) = 1$ and $\gamma(x) = \gamma(y) = \gamma(z) = 1$: their average, $\omega := 1/2\beta + 1/2\gamma$ has mixing entropy $S(\omega) = 1$. This follows from the fact that the only convex decomposition of ω into pure states is into β and γ, which in turn follows from the fact that these are the only pure states that assign probability one to z. On the other hand, $\omega(z) = 1$, so $H(\omega) = 0$.*

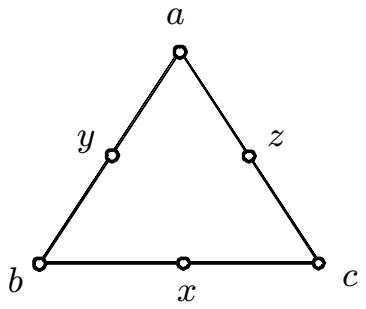

FIG. 3: The Greechie diagram for the test space of the firefly model.

Even in the general case, measurement entropy is quite well behaved. For example, it is easy to see that $H(\alpha)$ is continous as a function of α. Further,

Theorem 1 *Measurement entropy is concave, i.e., if $\sum_i t_i \alpha_i$ is a convex combination of states α_i on $\mathcal{A}$, then*

$$H\left(\sum_i t_i \alpha_i\right) \geq \sum_i t_i H\left(\alpha_i\right). \tag{1}$$

Proof. Since for each test E the local entropy H_E is concave,

$$\begin{aligned} H\left(\sum_i t_i \alpha_i\right) &= \inf_E H_E\left(\sum_i t_i \alpha_i\right) \\ &\geq \inf \sum_i t_i H_E(\alpha_i) \geq \sum_i t_i H(\alpha_i). \end{aligned}$$

□

Mixing entropy is, by contrast, a curious beast. The following example shows that it need not be continuous as a function of the state.

Example 2 *Let $\Omega \subseteq \mathbb{R}^3$ be the convex hull of the circle $C = \{(x, y, 0) | x^2 + y^2 = 1\}$ and the line segment $I =$*

$\{(1,0,t)|-1 \leq t \leq 1\}$ *(Figure 4). Let* α *denote the point of intersection of* I *and* C*, i.e., the point* $(1,0,0)$*. The extreme points of this set are evidently the endpoints of* I*, together with the points of* $C \setminus \{\alpha\}$*. Note that* α *has a unique decomposition as a mixture of extreme points of* Ω*, namely, as an equal mixture of the endpoints of* I*. Thus,* $S(\alpha) = 1$*. On the other hand,* α *can be approached as closely as we like by extreme points belonging to* $C \setminus \{\alpha\}$*, which have mixing entropy* 0*. The mixing entropy is therefore discontinuous at* α*.*

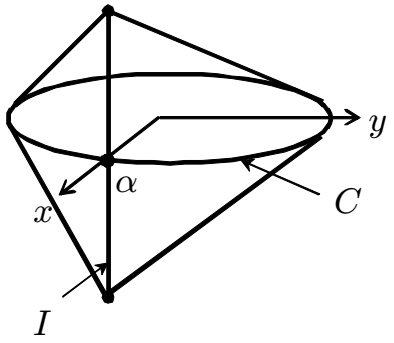

FIG. 4: Example of a state space Ω for which mixing entropy is not everywhere continuous (see Example 2).

Example 3 Let Ω be a square. Let α and β be the mid-points of adjacent faces, noting that these each have unit mixing entropy, $S(\alpha) = S(\beta) = 1$. Let $\gamma = 1/2(\alpha + \beta)$ be the mid-point of the line segment between α and β, and note that it also lies on the line segment between antipodal vertices of Ω (the diagonal through the square between the chosen faces). But given that γ is not at the midpoint of this diagonal, the Shannon entropy for the associated convex decomposition is less than one, as is therefore the infimum over convex decompositions. Therefore, the mixing entropy for γ satisfies $S(\gamma) < 1$. Consequently, $S(\gamma) < 1/2S(\alpha) + 1/2S(\beta)$, and we have a failure of concavity of the mixing entropy.

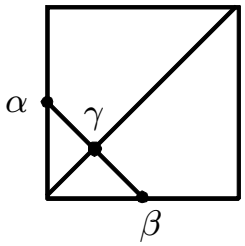

FIG. 5: Failure of concavity of mixing entropy for a squit.

In fact, the failure of concavity for the mixing entropy is quite generic.

Theorem 2 *Mixing entropy is not concave whenever the state space* Ω *is a non-simplicial polytope.*

The proof is given in Appendix A. It follows that an assumption of concavity for the mixing entropy forces the state space to be either a simplex (i.e. classical) or not a polytope. Hence such an assumption or one that implies it may be a useful tool in axiomatizing quantum theory.

It is natural to ask what follows from the condition that, as in classical and quantum theories, measurement and mixing entropy coincide. One immediate consequence is that mixing entropy will be concave. In view of Theorem 2, this implies that either the system is essentially classical, or there are an infinite number of pure states. Hence equality of measurement and mixing entropies narrows down possible theories quite a lot. We discuss this matter further in Appendix B.

Both measurement and mixing entropy have been considered before, notably by Hein [20], in a similar context, albeit with somewhat different aims than ours in view. There are various other entropic quantities one could reasonably consider. For example, a concept of entropy that might be more closely related to operational tasks is the supremum, over convex decompositions of the state and over tests, of the classical mutual information between the random variable specifying the element of the convex decomposition, and the random outcome of the test. Natural analogues of this quantity and of the measurement and preparation entropies defined above exist in the closely related ordered linear spaces framework (also known as the convex sets framework) for theories. Test space models such as we have defined above induce ordered linear spaces models by a linearization procedure that embeds the test space in a vector space and identifies outcomes in the test space with certain elements of the dual vector space; this procedure allows one to define concepts of measurement entropy more tightly related to the geometry of the state space, but that can usually be viewed as special cases of the test space definition. Appendix C gives a further brief discussion of this.

From this point on, we focus mainly on measurement entropy. As always with mathematical definitions, there is a certain tension between the ideals of flexibility and generality, on the one hand, and, on the other, the desire to avoid annoying pathologies. Our test-space dependent definition of measurement entropy definitely errs on the side of the former, in that it is consistent with quite absurd examples. For example, if one includes in one's test space a test having a single outcome, then all states will automatically have zero entropy. One can avoid such difficulties by placing various restrictions on the test spaces to be considered, at the cost of a slightly more involved technical development. Going to the linearized setting mentioned above may also help. Our work in this paper

does not demand such fastidiousness, however, as our results are of a very general character.

IV. COMPOSITE SYSTEMS AND JOINT ENTROPY

Most of the interesting problems of information theory involve more than one system. The following subsection describes how to treat composite systems in the language of test spaces. The idea is that, given systems A and B, the joint system AB should be associated with a test space and state space of its own. However, there is not a unique recipe for determining test and state spaces for AB given the test and state spaces for A and B. Instead, a theory must give additional rules that specify how systems combine.[46] Our results will pertain to a variety of notions of composition, although we limit the scope by requiring certain properties to hold. In particular, we assume that the test space of the composite includes all product tests and conditional two-stage tests (where one party's choice of test is conditioned on the outcome of the other party's test). One motivation for this is to have a test space that is sufficiently rich to be interesting. Another motivation is that this assumption guarantees that all states are non-signaling. We go on to define analogues of familiar quantities, such as joint entropies and the mutual information, which are used later to analyze information causality.

A. Composite Systems

Consider two systems, A and B, where $A = (\mathfrak{A}, \Omega^A)$ and $B = (\mathfrak{B}, \Omega^B)$. For convenience, assume that these are controlled by two parties, called Alice and Bob. The first, and most basic, assumption we shall make is that Alice can perform any test $E \in \mathfrak{A}$ simultaneously with Bob performing any test $F \in \mathfrak{B}$. This can be regarded as a single *product test*. The possible outcomes of this product test are pairs of the form $(e, f) \in E \times F$.

Definition 3 *The* Cartesian product *of the test spaces* $\mathfrak{A}$ *and* $\mathfrak{B}$ *is the collection of all product tests. It is denoted* $\mathfrak{A} \times \mathfrak{B}$.

The set $\Omega(\mathfrak{A} \times \mathfrak{B})$, of all states that can be defined on the Cartesian product test space, typically includes *signaling* states, which allow Alice to send messages instantaneously to Bob, or vice versa, by varying her choice of which test to perform.

Definition 4 *A state* ω^{AB} *on* $\mathfrak{A} \times \mathfrak{B}$ *is* non-signaling *iff*

$$\begin{aligned} \sum_{f \in F} \omega^{AB}(e, f) &= \sum_{f' \in F'} \omega^{AB}(e, f') \quad \forall e, F, F' \\ \sum_{e \in E} \omega^{AB}(e, f) &= \sum_{e' \in E'} \omega^{AB}(e', f) \quad \forall f, E, E'. \end{aligned} \tag{2}$$

If a state ω^{AB} is non-signaling, it is possible to define the *marginal* (or *reduced*) state ω^A via

$$\omega^A(e) = \sum_{f \in F} \omega(e, f), \tag{3}$$

where Eq. (2) ensures that the right hand side is independent of $F \in \mathfrak{B}$. The marginal ω^B is defined similarly.

If ω^{AB} is non-signaling, it is also possible to define a *conditional state*, $\omega^{B|e}$. Informally, this is the updated state at Bob's end following the outcome e being obtained for a test at Alice's end:

$$\omega^{B|e}(f) := \omega^{AB}(e, f)/\omega^A(e).$$

By convention, $\omega^{B|e}$ is zero if $\omega^A(e)$ is zero. The conditional state $\omega^{A|f}$ is defined similarly.

Notice that a particular type of measurement, which might be thought reasonable, is not included in the Cartesian product. This is a joint measurement, where Alice first measures her system, and communicates the result to Bob, who performs a measurement which depends on Alice's outcome. Entangled measurements, such as are allowed in quantum theory, are also not included. Hence the Cartesian product $\mathfrak{A} \times \mathfrak{B}$ models a situation in which Alice and Bob are fairly limited - they can act independently and collate the results of their actions at a later time, but cannot otherwise communicate.

It is possible to construct a more sophisticated product of two test spaces, which does allow for the kind of two stage measurements just described (although still not entangled measurements). Let $\overleftrightarrow{\mathfrak{A}\mathfrak{B}}$ denote the test space consisting of the following

1. All two-stage tests, where a test $E \in \mathfrak{A}$ is performed, and then, depending on the outcome e that is obtained, a pre-selected test $F_e \in \mathfrak{B}$ is performed.
2. All two-stage tests, where a test $F \in \mathfrak{B}$ is performed, and then, depending on the outcome f that is obtained, a pre-selected test $E_f \in \mathfrak{A}$ is performed.

$\overleftrightarrow{\mathfrak{A}\mathfrak{B}}$ is called the *Foulis-Randall* or *bilateral* product of test spaces $\mathfrak{A}$ and $\mathfrak{B}$.

The Foulis-Randall product contains the Cartesian product, $\mathfrak{A} \times \mathfrak{B} \subseteq \overleftrightarrow{\mathfrak{A}\mathfrak{B}}$, because product tests are a special case of two-stage tests. Furthermore, if either A or B is non-classical, then not all two-stage tests are product tests, so that the containment is strict. The containment of one test space in another has consequences for their state spaces. Specifically, if $\mathfrak{X}$ and $\mathfrak{Y}$ are test spaces such that $\mathfrak{X} \subseteq \mathfrak{Y}$, then the convex set $\Omega(\mathfrak{Y})$ may be in a higher dimensional space than $\Omega(\mathfrak{X})$, but the restrictions of states in $\Omega(\mathfrak{Y})$ to $\mathfrak{X}$ (which are well-defined, since every test in $\mathfrak{X}$ is also a test in $\mathfrak{Y}$) are all contained in $\Omega(\mathfrak{X})$. In other words, writing $\Omega(\mathfrak{Y})|_{\mathfrak{X}}$ for the set of restrictions to $\mathfrak{X}$ of states on $\mathfrak{Y}$, we have $\Omega(\mathfrak{Y})|_{\mathfrak{X}} \subseteq \Omega(\mathfrak{X})$. Because the additional measurements in $\mathfrak{Y}$ place additional constraints on these states, the containment may well be strict.

It follows that the restriction of the maximal state space of the Foulis-Randall product to the Cartesian product is contained within the maximal state space of the Cartesian product, $\Omega(\overleftrightarrow{\mathfrak{A}\mathfrak{B}})|_{\mathfrak{A}\times\mathfrak{B}} \subseteq \Omega(\mathfrak{A} \times \mathfrak{B})$. The containment is strict if one of the systems is non-classical. Indeed, the states in $\Omega(\overleftrightarrow{\mathfrak{A}\mathfrak{B}})|_{\mathfrak{A}\times\mathfrak{B}}$ correspond exactly to the non-signaling states in $\Omega(\mathfrak{A} \times \mathfrak{B})$. This is demonstrated in Ref. [16].

We are now prepared to define the class of test and state spaces for composites in which we shall be interested. The test space for the composite, which we denote by $\mathfrak{C}$, is required to contain the Foulis-Randall product of the components, $\overleftrightarrow{\mathfrak{A}\mathfrak{B}} \subseteq \mathfrak{C}$. The state space of the composite, which we denote by Ω^{AB}, is unconstrained beyond being a subset of the maximal state space, $\Omega^{AB} \subseteq \Omega(\mathfrak{C})$. Recalling that $\overleftrightarrow{\mathfrak{A}\mathfrak{B}} \subseteq \mathfrak{C}$ implies $\Omega(\mathfrak{C})|_{\overleftrightarrow{\mathfrak{A}\mathfrak{B}}} \subseteq \Omega(\overleftrightarrow{\mathfrak{A}\mathfrak{B}})$ and that all the states in $\Omega(\overleftrightarrow{\mathfrak{A}\mathfrak{B}})$ are non-signalling, it follows that all states in Ω^{AB} are non-signalling. Indeed, the main motivation for confining our attention to test spaces containing $\overleftrightarrow{\mathfrak{A}\mathfrak{B}}$ is that this is sufficient to ensure no-signalling without any further constraints on the state space.

Given a state $\omega^{AB} \in \Omega^{AB}$, the marginals ω^A, ω^B, and conditionals of the form $\omega^{A|f}$, $\omega^{B|e}$, are defined in the obvious way by the probabilities which ω^{AB} assigns to the product tests. Furthermore, we assume that the composite systems we consider satisfy the following natural requirement: that if a test is performed on system A, the conditional state on system B must be allowed in the theory, i.e., be contained in Ω^B, and vice versa. Hence Ω^{AB} satisfies the constraint that for all e and f such that $\omega^A(e), \omega^B(f) \neq 0$, $\omega^{B|e}$ and $\omega^{A|f}$ belong to Ω^B and Ω^A respectively. This is enough to ensure that the marginal states ω^A, ω^B also belong to the state spaces of the component systems.

A general composite test space $\mathfrak{C}$ may contain non-product measurements, which are not contained in the Foulis-Randall product. Quantum theory, for instance, has a test space for composites that is larger than the Foulis-Randall product. If A and B are quantum systems, so that $A = (\mathfrak{F}(\mathbf{H}), \Omega(\mathbf{H}))$ and $B = (\mathfrak{F}(\mathbf{K}), \Omega(\mathbf{K}))$, then the quantum joint system is $AB := (\mathfrak{F}(\mathbf{H} \otimes \mathbf{K}), \Omega(\mathbf{H} \otimes \mathbf{K}))$, which is a composite in our sense and contains non-product measurement outcomes, for instance, entangled ones.

Henceforth, AB will stand for a general non-signaling composite of systems A and B. In the particular case where $A = (\{E\}, \Delta(E))$ is a classical system, we always take $\mathfrak{C}$ to be the Foulis-Randall product $\overleftrightarrow{\{E\}\mathfrak{B}}$. We also assume that composition of systems is associative, so that for any three systems A, B and C, there is a natural isomorphism $A(BC) \simeq (AB)C$.

In addition to specifying how systems combine, a probabilistic theory must specify what sorts of systems are allowed. For instance, in finite-dimensional quantum theory, every dimensionality of Hilbert space defines a different type of system and they are all allowed. Furthermore, a classical system of arbitrary dimensionality (that is, arbitrary cardinality for the test) can be defined within quantum theory as a restriction upon a quantum system of the same dimensionality, so in this sense classical systems are allowed as well. A probabilistic theory must specify the types of systems that are allowed and how these compose. We shall confine our attention to theories incorporating only finite-dimensional systems, and those that contain, for any finite set E, the classical system $(E, \Delta(E))$. (Thus, for us, *quantum theory* means finite-dimensional quantum theory in conjunction with classical systems.) For a discussion of what such theories might look like in category-theoretic terms, see [9, 10].

B. Joint Entropies, Conditional Entropies, Mutual Information

Consider a composite system $AB = (\mathfrak{C}, \Omega^{AB})$. The measurement entropy $H(\omega^{AB})$ of a state $\omega^{AB} \in \Omega^{AB}$, which we will sometimes denote by $H(AB)$, is the infimum over $E \in \mathfrak{C}$ of $H_E(\omega^{AB})$. In this context, it will also be understood that $H(A)$ and $H(B)$ stand for the entropies $H(\omega^A)$ and $H(\omega^B)$ of the marginal states ω^A and ω^B.

Theorem 3 *Measurement entropy is subadditive. That is, for any composite AB,*

$$H(AB) \leq H(A) + H(B).$$

Proof. Let ω be the joint state of AB, with marginal states ω^A and ω^B. Choose E and F with $H(\omega^A) = H_E(\omega^A)$ and $H(\omega^B) = H_F(\omega^B)$. By the definition of measurement entropy, the definition of a composite, and the subadditivity of Shannon entropy, we have $H(AB) \leq H_{EF}(\omega) \leq H_E(\omega^A) + H_F(\omega^B) = H(A) + H(B)$. □

Definition 5 *The* conditional measurement entropy *between A and B is defined to be*

$$H(A|B) := H(AB) - H(B). \tag{4}$$

Our notation here is less precise than it might be, since the joint entropy $H(AB)$ depends on the test space associated with the joint system, hence so do conditional entropies. We will try to be clear, at any point where the question could arise, as to what product is in play.

Classically, given a joint distribution ω^{AB} over variables A and B, one defines the *mutual information* by

$$I(A:B) = H(A) + H(B) - H(AB), \tag{5}$$

where H denotes the Shannon entropy. One can regard this as a measure of how far A and B are from being independent: by subadditivity, $I(A:B) \geq 0$, with $I(A:B) = 0$ iff A and B are independent, i.e., ω^{AB} factorizes. In attempting to extend the concept of mutual information to more general models, one might very naturally consider defining $I(A:B)$ to be the maximum of the mutual informations $I(E:F)$ as E and F range over tests belonging to systems A and B, respectively. However, the usual practice in quantum theory is simply to take Equation (5), with von Neumann entropies replacing Shannon entropies, as *defining* mutual information. In general, this gives a different value. In order to facilitate comparison with quantum theory, we shall adopt the following

Definition 6 *Let AB be a composite system. The* measurement-entropy-based mutual information *between A and B is*

$$I(A:B) := H(A) + H(B) - H(AB). \tag{6}$$

With this definition, the subadditivity of measurement entropy (Theorem 3) implies that measurement-entropy-based mutual information is non-negative. Hereafter, we will refer to this simply as the "mutual information". Note that Eq. (5) is a special case of this definition.

Now intuitively, one might expect that the mutual information $I(A:B)$ between two systems should not *decrease* if we recognize that B is a part of some larger composite system BC – i.e., that $I(A:B) \leq I(A:BC)$. Simple algebraic manipulations (using Eqs. (4) and (6)) allow us to reformulate this condition in various ways.

Lemma 1 *The following are equivalent:*

(a) $I(A:BC) \geq I(A:B)$

(b) $H(A|BC) \leq H(A|B)$

(c) $H(A,B) + H(B,C) - H(B) \leq H(A,B,C)$

(d) $I(A:B|C) \geq 0$, *where* $I(A:B|C) = H(A|C) + H(B|C) - H(AB|C)$.

Definition 7 *The measurement entropy is said to be* strongly subadditive *if it satisfies the equivalent conditions (a)-(d).*

(We use this terminology despite the fact that it is usually only condition (c) that goes by the name of "strong subadditivity" and despite the fact that conditions (a) and (d) constrain the measurement entropy only through the definitions of $I(A:BC)$ and $I(A:B|C)$.) A probabilistic theory in which conditions (a)-(d) are satisfied for all systems A, B and C will also be called *strongly subadditive*.

Both the Shannon and von Neumann entropies are strongly subadditive. In the former case, this is a straightforward exercise, but in the latter, a relatively deep fact. Colloquially, this means that in classical and quantum theories, just forgetting about or discarding a system C never increases one's mutual information between systems A and B. As the following shows, however, strong subadditivity can fail in general theories, even when two of the three systems are classical. One potential gloss is that discarding or forgetting about system C *can* increase the mutual information between systems A and B. But a more sensible reading is perhaps that the quantity defined as mutual information should not in the general case be interpreted as "the information one system contains about another."

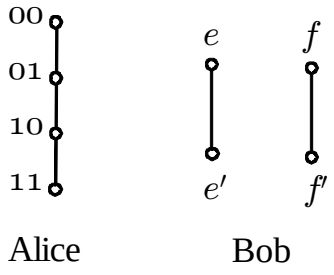

FIG. 6: The component test spaces for example 4.

Example 4 *(Failure of strong subadditivity of the measurement entropy.) Consider a tripartite system ABC, where A and B are classical bits and C is a squit, with*

test space $\{\{e,e'\},\{f,f'\}\}$. *Consider the joint state described by the following table:*

	e	e'	f	f'
00	1/4	0	1/4	0
01	1/4	0	0	1/4
10	0	1/4	1/4	0
11	0	1/4	0	1/4

In words, the outcome of test $\{e,e'\}$ *is perfectly correlated with* A *while the outcome of the test* $\{f,f'\}$ *is perfectly correlated with* B. *It is easily verified that*

$$H(C) = H(AC) = H(BC) = 1.$$

If all three systems are measured, with the test on C *perhaps depending on the values of* A *and* B, *there are always four distinct outcomes, each with probability* 1/4. *Hence* $H(ABC) = 2$ *and*

$$\begin{aligned} I(A:B|C) &= H(AC) + H(BC) - H(C) - H(ABC) \\ &= 1 + 1 - 1 - 2 = -1 < 0, \end{aligned}$$

which contradicts form (d) of strong subadditivity.

Note that the foregoing example is all but classical, depending not on any notion of entanglement or nonlocality, but only on the fact that one can measure either, but never both, of $\{e,e'\}$ and $\{f,f'\}$.

This section concludes with some lemmas, which hold in the special case that one or more of the systems in the composite is classical. Some of these are useful later on.

Lemma 2 *Let* ω^{AB} *be a state on* AB, *where* A *is classical. Then*

$$H(\omega^{AB}) = H(\omega^A) + \sum_{e\in E} \omega^A(e) H(\omega^{B|e}). \tag{7}$$

The proof is straightforward. As a shorthand, when A is classical we might write

$$H(AB) = H(A) + \sum_e p(e) H(B|e).$$

Corollary 1 *If* A *is classical, then* $H(B|A) \geq 0$ *for any system* B.

The proof is immediate from Eqs. (4) and (7).

Corollary 2 *If* A *is classical and independent of* B, *then* $H(AB) = H(A) + H(B)$.

Proof. The assertion that A and B are independent means that the joint state is $\omega^{AB} = \omega^A \otimes \omega^B$, i.e., that $\omega^{B|e} = \omega^B$ for all $e \in E$. By Lemma 2, we have

$$\begin{aligned} H(AB) &= H(A) + \sum_{e\in E} \omega^A(e) H(\omega^{B|e}) \\ &= H(A) + \left(\sum_{e\in E} \omega^A(e)\right) H(B) = H(A) + H(B). \end{aligned}$$

□

Finally, strong subadditivity does hold in the special case that systems A and C in Lemma 1 are classical. Colloquially, discarding a *classical* system can never result in an increase in the mutual information between a general system and another classical system.

Lemma 3 *Let* A *and* C *be classical. Then for any system* B,

$$H(A|BC) \leq H(A|B).$$

Hence, the equivalent conditions of Lemma 1 are satisfied.

Proof. Let $A = \{E\}$, $C = \{G\}$, and let the joint state of ABC be ω^{ABC}. Then the marginal state of BC satisfies $\omega^{BC}(fg) = \omega^C(g) \cdot \omega^{B|g}(f)$ where, for all $g \in G$

$$\omega^{B|g} = \sum_{e\in E} \frac{\omega^{AC}(eg)}{\omega^C(g)} \omega^{B|eg}.$$

By Lemma 2, we have

$$\begin{aligned} H(A|BC) &= H(ABC) - H(BC) \\ &= H(AB) + \sum_{g\in G}\sum_{e\in E} \omega^{AC}(eg) H(\omega^{B|eg}) \\ &\quad - H(C) - \sum_{g\in G} \omega^C(g) H(\omega^{B|g}). \end{aligned}$$

We can rewrite this as

$$\begin{aligned} H(A|BC) &= H(A|C) \\ &\quad + \sum_{g\in G}\sum_{e\in E} \omega^{AC}(eg) H(\omega^{B|eg}) \\ &\quad - \sum_y \omega^C(g) H(\omega^{C|g}). \end{aligned} \tag{8}$$

Since measurement entropy is concave,

$$H(\omega^{C|g}) \geq \sum_{e\in E} \frac{\omega^{AC}(eg)}{\omega^C(g)} H(\omega^{B|g}),$$

whence

$$\sum_g \omega^C(g) H(\omega^{B|g}) \geq \sum_{g\in G}\sum_{e\in E} \omega^{AC}(eg)\omega^{C|eg}.$$

It follows that $\sum_{e,g} \omega^{AC}(eg) H(\omega^{B|eg}) - \sum_y \omega^C(g) H(\omega^{B|g}) \leq 0$, which, combined with Equation (8), gives the desired result that $H(A|BC) \leq H(A|C)$. □

C. Data Processing and the Holevo Bound

A fundamental result in quantum information theory, the *Holevo bound*, asserts that if Alice prepares a quantum state $\rho = \sum_{x \in E} p_x \rho_x$ for Bob, then, for any measurement F that Bob can make on his system,

$$I(E:F) \leq \chi,$$

where $\chi := H(\rho) - \sum_{x \in E} p_x H(\rho_x)$ (often called the *Holevo quantity*).

This inequality makes sense in our more general setting. Suppose that Alice has a classical system $A = (\{E\}, \Delta(E))$ and Bob a general system B. Alice's system is to serve as a record of which state of B she prepared. Hence the situation above is modeled by the joint state $\omega^{AB} = \sum_{x \in E} p_x \delta_x \otimes \beta_x$, where δ_x is a deterministic state of Alice's system with $\delta_x(x) = 1$. Bob's marginal state is $\omega^B = \sum_{x \in E} p_x \beta_x$. By Lemma 2, $H(\omega^{AB}) = H(A) + \sum_{x \in E} p_x H(\beta_x)$. Hence,

$$\begin{aligned} I(A:B) &= H(A) + H(B) - H(AB) \\ &= H(A) + H(B) - \left(H(A) + \sum_{x \in E} p_x H(\beta_x) \right) \\ &= H(\omega^B) - \sum_{x \in E} p_x H(\beta_x) = \chi. \end{aligned}$$

Accordingly, the content of the Holevo bound is simply that the mutual information between the measurement of Alice's classical system and any measurement on Bob's system is no greater than $I(A:B)$,

$$I(E:F) \leq I(A:B).$$

While this is certainly natural, it does not always hold.

Example 5 *(Failure of the Holevo bound.) Let A be a classical bit, $A = \{\{0,1\}\}$ and B a squit, $B = \{F = \{f, f'\}, G = \{g, g'\}\}$, and consider the state*

	f	f'	g	g'
0	1/2	0	1/2	0
1	1/2	0	0	1/2

It is easy to check that $H(A) = H(AB) = H_G(B) = 1$, and $H(B) = H_F(B) = 0$. Hence,

$$I(A:B) = H(A) + H(B) - H(AB) = 1 + 0 - 1 = 0$$

while

$$I(E:F) = 1 + 1 - 1 = 1 > 0.$$

Both strong subadditivity and the Holevo bound are instances of a more basic principle. The *data processing inequality* (DPI) asserts that, for any systems A and B and any physical process $\mathcal{E} : B \to C$,[47]

$$I(A : \mathcal{E}(B)) \leq I(A:B).$$

The strong subadditivity of entropy amounts to the DPI for the process that simply discards a system (the *marginalization map* $BC \to C$). The Holevo bound is the DPI for the special case of measurements, which can be understood as processes taking a system into a classical system which records the outcome.

It seems reasonable that discarding a system, or performing a measurement, should be allowed processes in a physical theory. But a notion of mutual information, according to which discarding a system, or performing a measurement, causes a *gain* of mutual information seems bizarre. So it is an attractive idea that a physical theory should allow at least some definition of entropy and mutual information such that the corresponding DPI is satisfied.

V. INFORMATION CAUSALITY

In [27], Pawlowski *et al.* define a principle they call *Information Causality* in terms of the following protocol. Alice and Bob share a joint non-signaling state, known to both parties. Alice receives a random bit string E of length N, makes measurements, and sends Bob a message F of no more than m bits. Bob receives a random variable G encoding a number $k = 1, ..., N$, instructing him to guess the value of Alice's kth bit E_k. Bob thereupon makes a suitable measurement and, based upon its outcome, and the message from Alice, produces his guess, b_k. Information causality is the condition that

$$\sum_{k=1}^{N} I(E_k : b_k | G = k) \leq m. \tag{9}$$

The main result of [27] is that if a theory contains states that violate the CHSH inequality [11] by more than the Tsirel'son bound [33], then it violates information causality. In particular, if Alice and Bob can share PR boxes, then using a protocol due to van Dam [34], they can violate information causality maximally, meaning that Bob's guess is correct with certainty, and the left hand side of Equation (9) is N. Pawlowski *et al.* also give a proof, using fairly standard manipulations of

quantum mutual information, that quantum theory *does* satisfy information causality.

Having seen how to define notions of entropy and mutual information for general systems, it is interesting to consider where Pawlowski *et al.*'s quantum proof breaks down for some non-quantum systems such as PR boxes. One issue is that the proof uses strong subadditivity. As the following subsection shows, in the case where a PR box is the shared state, the van Dam protocol itself provides an example of the failure of strong subadditivity of the measurement entropy. Section V B provides a converse result. Any theory which is monoentropic, strongly subadditive and where the Holevo bound holds, must satisfy information causality.

First, a few words about how to describe this setting in our terminology. Let Alice and Bob share two systems A and B, where each of these, as usual, has an associated test space. The joint test space of AB is immaterial, as long as it includes the Foulis-Randall product (i.e., allows all the separable measurements). The bit strings E and F are regarded as classical systems in their own right and the joint test space for a classical and a general system is, as always, assumed to be the Foulis-Randall product. Systems A and B begin the protocol in some joint non-signaling state ω^{AB}.

A. The van Dam protocol.

Consider a special case of the protocol described above, in which Alice and Bob share a PR box. Alice is supplied with a two-bit string $E = E_1E_2$, and transmits one bit F to Bob. Let the PR box be a state of two systems A and B, where A and B are squits corresponding to the test spaces $\{\{a_1, a_1'\}, \{a_2, a_2'\}\}$ and $\{\{b_1, b_1'\}, \{b_2, b_2'\}\}$ respectively. The joint state of A and B is

	a_1	a_1'	a_2	a_2'
b_1		1/2		1/2
b_1'	1/2		1/2	
b_2		1/2	1/2	
b_2'	1/2			1/2

It can be verified that these outcome probabilities are indeed the PR box correlations, violating the CHSH inequality maximally. In van Dam's protocol, Alice determines the parity, $E_1 \oplus E_2$ (where $\oplus$ denotes addition mod 2). If this is zero she performs the $\{a_1, a_1'\}$ measurement on her system; if it is 1, she performs the $\{a_2, a_2'\}$ measurement. She then sends Bob a single bit with a value equal to the parity of her outcome and E_1 (where unprimed outcomes correspond to 0 and primed outcomes to 1). Bob can then determine the value of E_1 by measuring $\{b_1, b_1'\}$, or the value of E_2 by measuring $\{b_2, b_2'\}$.

Consider now an intermediate stage in this protocol, at which Alice has measured her system, and sent the bit F to Bob, who has not yet measured his system. Bob has access to systems B and F, but does not know the outcome of Alice's measurement. Hence consider the joint state of EFB, averaged over the outcomes of Alice's measurement. This is easily verified to be

E_1E_2F	b_1	b_1'	b_2	b_2'
000	1/8		1/8	
001		1/8		1/8
111	1/8		1/8	
110		1/8		1/8
010	1/8			1/8
011		1/8	1/8	
101	1/8			1/8
100		1/8	1/8	

Minimizing over the possible measurement choices on B,

$$H(E_1, F, B) = H(E_2, F, B) = H(F, B) = 2.$$

But clearly, $H(E, B) = 3$, so

$$I(E_1 : E_2|F, B) = 2 + 2 - 2 - 3 = -1 < 0.$$

B. Theories satisfying information causality

As the previous subsection observes, the van Dam protocol involves a joint state on a classical-nonclassical composite system, which does not satisfy strong subadditivity of entropy. This is enough to prevent the proof of information causality going through. This subsection proves a converse result.

Theorem 4 *Suppose that a theory is*

1. *monoentropic, meaning that measurement entropy equals mixing entropy for all systems.*
2. *Strongly subadditive*
3. *Satisfies the Holevo bound.*

Then the theory satisfies information causality. It follows that any theory satisfying these conditions cannot violate Tsirel'son's bound.

Note that, as discussed in Section IV C, the second and third conditions both follow from a single assumption of a data processing inequality. Note also that in proving Theorem 4, the condition that a theory be monoentropic is only used to establish the technical condition

that $H(A|B) \geq 0$ when A is classical.[48] So the theorem would still be valid if the monoentropic assumption were replaced by a direct assumption that for classical A, $H(A|B) \geq 0$. Otherwise, begin with

Lemma 4 *Suppose that a theory is monoentropic and that A is a classical system. Then $H(A|B) \geq 0$ for any system B.*

Proof. (Lemma 4.) Suppose that A is a classical system, and that the joint state of AB is ω^{AB}. If the measurement and mixing entropies are equal, then Lemma 2 immediately gives

$$S(AB) = S(A) + \sum_x p_x S(\beta_x),$$

where $p_x = \omega^A(x)$ and β_x is the state of B conditioned on x. Recall that the mixing entropy of a state is defined in terms of an infimum over convex decompositions into pure states. For a fixed ϵ, call a convex decomposition of a state ω into pure states ϵ-*optimal* if the Shannon entropy of the coefficients is $\leq S(\omega) + \epsilon$. For any $\epsilon > 0$, there is an ϵ-optimal decomposition. Let

$$\beta_x = \sum_y q_{y|x} \beta_{xy}$$

be an ϵ-optimal convex decomposition of β_x into pure states β_{xy}. It follows that

$$\omega^B = \sum_x \sum_y p_x q_{y|x} \beta_{xy}$$

is a (possibly far from optimal) convex decomposition of ω^B into pure states. Hence $S(B)$ is less than or equal to the Shannon entropy of the coefficients on the right hand side. Therefore

$$\begin{aligned} S(B) &\leq H(p_x) + \sum_x p_x H(q_{y|x}) \\ &\leq S(A) + \sum_x p_x (S(\beta_x) + \epsilon) \\ &= S(AB) + \epsilon. \end{aligned}$$

Since this holds for any ϵ, we have $S(B) \leq S(AB)$ and $S(A|B) = S(AB) - S(B) \geq 0$ as required. □

Given Lemma 4, the proof of Theorem 4 is essentially a reconstruction of the quantum argument of Appendix A of [27], adapted to the broader setting of non-signaling states on test spaces. In its form the proof is the same, but great care must be taken at each step to ensure that the relevant properties of entropies and mutual information still hold. Many of the steps still go through in virtue of generic properties of the measurement entropy. The explicit assumptions of Theorem 4 are needed for the rest.

Proof. (Theorem 4.) Assume that Alice and Bob share a joint system AB. Consider the N-bit string which Alice receives as a classical system E, and consider the m-bit message which Alice sends to Bob as a classical system F. Let E_k denote Alice's kth bit. Consider the stage of the protocol where Alice has measured system A, and sent F to Bob, but Bob has not yet measured system B. Bob has control of systems F and B at this point, and does not know the outcome of Alice's measurement. Hence the strategy is to consider the joint state of the systems E, F and B, averaged over Alice's outcomes.

The first goal is to show that the joint state at this point satisfies

$$I(E : FB) \leq m. \tag{10}$$

By the fact that the initial state of AB is non-signaling, E is independent of B. Therefore Corollary 2 yields $I(E : B) = 0$. Using this, along with the definitions and straightforward algebraic manipulation, we get

$$\begin{aligned} I(E : FB) &= I(E : B) + I(E : F|B) \\ &= I(E : F|B) \\ &= I(EB : F) - I(B : F). \end{aligned}$$

By Theorem 3, mutual information is non-negative, so

$$I(E : FB) \leq I(EB : F). \tag{11}$$

Now, $I(EB : F) = H(EB) + H(F) - H(EFB) = H(F) - H(F|EB)$. By the assumption that the theory is monoentropic, and Lemma 4, $H(F|EB) \geq 0$. So

$$I(EB : F) \leq H(F) \leq m. \tag{12}$$

This gives Equation (10).

The next step is to establish

$$\sum_{k=1}^{N} I(E_k : FB) \leq I(E : FB). \tag{13}$$

Rearrangement of definitions yields

$$\begin{aligned} I(E : FB) &= I(E_1 \ldots E_N : FB) \\ &= I(E_1 : FB) + I(E_2 \ldots E_N : FB|E_1) \end{aligned}$$

and

$$\begin{aligned} &I(E_2 \ldots E_N : FB|E_1) = \\ &\qquad I(E_2 \ldots E_N : FBE_1) - I(E_2 \ldots E_N : E_1). \end{aligned}$$

Since the distribution on E is uniform (the bits are independent), $I(E_2 \ldots E_N : E_1) = 0$. Hence,

$$I(E : FB) = I(E_1 : FB) + I(E_2 \ldots E_N : FBE_1).$$

By strong subadditivity,

$$I(E_2 \ldots E_N : FBE_1) \geq I(E_2 \ldots E_N : FB).$$

So

$$I(E : FB) \leq I(E_1 : FB) + I(E_2 \ldots E_N : FB).$$

Applying this inequality recursively gives Equation (13).

Finally, consider the last stage of the protocol. If Bob is instructed to guess the kth bit, then, depending on the message F, he measures system B. This can be seen as a single joint measurement X_k on the system FB. [49] The Holevo bound, combined with Equations (10,13) gives

$$\sum_{k=1}^{N} I(E_k : X_k) \leq m.$$

Finally, Bob outputs a guess b_k for the value of E_k, where the guess depends on k and on the outcome of the measurement X_k. The usual data processing inequality applied to classical mutual information yields

$$\sum_{k=1}^{N} I(E_k : b_k | G = k) \leq m,$$

which is information causality. □

VI. CONCLUSIONS, DISCUSSION AND FURTHER QUESTIONS

We have defined preparation and measurement based generalizations of quantum and classical entropy and mutual and conditional information, and studied some of their basic properties. We called theories in which they coincide *monoentropic*, and showed that if they in addition satisfy the data processing inequality (or at least its corollaries strong subadditivity and the generalized Holevo bound), Pawlowski *et al.*'s information causality principle holds. By their remarkable result that any correlations violating the Tsirel'son bound can be used to violate information causality, it follows that monoentropic theories satisfying data processing must, like quantum theory, obey the Tsirel'son bound. Monoentropicity, is a strong constraint on theories, as we have shown by establishing that it fails for all polytopes except simplices.

Our results indicate that it is interesting and profitable to develop notions of entropy, and allied notions of conditional entropy and mutual information, for abstract probabilistic models. This paper should be regarded as only a preliminary exploration of this possibility.

A natural direction for further research is to study data compression and channel capacities in the abstract setting of this paper. It is natural to seek a measure of entropy that governs the rate of high-fidelity data compression, as Shannon and von Neumann entropy do in classical and quantum theory. A first step toward exploring *classical* channel capacities in generalized probabilistic theories might be to identify sufficient conditions for the Holevo bound to hold. This is related to the issue of finding an operationally motivated definition of mutual information. Arguably, a properly motivated notion of mutual information should *manifestly* be monotonic. Of course, the monotonicity of quantum mutual information—equivalently, the strong subadditivity of quantum entropy—is not manifest from its usual functional form. Still, the outright failure of the measurement-entropy-based mutual information to satisfy monotonicity in some cases raises a question as to its significance. Although in such cases measurement-based mutual information cannot be used to establish information causality through a proof parallel to Pawlowski *et al.*'s quantum proof, it could be that IC nevertheless holds in some such cases. One should be cautious, though, about dismissing natural generalizations of classical quantities on the grounds that they fail to satisfy intuitively compelling properties. A case in point is the history of skepticism, based on the fact that it can be negative, about the operational significance of conditional information in quantum information theory. It was known for many years that the conditional mutual information can be negative, but it was eventually shown to have an operational interpretation, involving the rate for quantum state merging protocols. It is also good to keep in mind that different operational motivations might turn out to be naturally associated with different entropic quantities, each with reasonable claim to be called mutual information.

At a more fundamental level, one would like to understand better the operational significance of various notions of entropy for abstract probabilistic models and theories. It is likely that the entropic quantities we have discussed here, measurement and mixing entropy, will turn out not to be best notions of entropy to use in many situations. For example, in Appendix C, we considered a variation (or perhaps better, a specialization) of the notion of measurement entropy that is more tightly coupled to the geometry of the state space.

We have seen that, taken together, the conditions of monoentropicity, strong subadditivity, and the Holevo bound, imply information causality. It is not out of the question that some subset of these conditions would suffice (especially since we need only very special cases of strong subadditivity). Alternatively, it would be of interest to find a single, reasonably simple physical postulate that would imply all three of these conditions. It seems plausible that such a postulate exists. On the one hand, strong subadditivity and the Holevo bound are both special cases of the data processing inequality, which in turn can be derived (as we will detail in a future paper) from the assumption that arbitrary processes can be dilated to reversible ones. On the other hand, as we show in

Appendix B, monoentropicity can be derived from conditions of a similar flavor, involving the dilatability of mixed states to pure states with a "marginal steering" property. Another avenue to explore is the consequence of monoentropicity that is needed for the IC proof: positivity of conditional information when a classical system is conditioned upon a general one. Although its operational interpretation is not evident at first blush, it warrants further study.

We hope to discuss all of these matters in detail in a future paper.

Acknowledgments

This research was supported by the United States Government through grant OUR-0754079 from the National Science Foundation. It was also supported by Perimeter Institute for Theoretical Physics. Research at Perimeter Institute is supported by the Government of Canada through Industry Canada and by the Province of Ontario through the Ministry of Research and Innovation. This work was also supported by the EU's FP6-FET Integrated Projects SCALA (CT-015714) and QAP (CT-015848), and the UK EPSRC project QIP-IRC. Jonathan Barrett is supported by an EPSRC Career Acceleration Fellowship. At IQC, Matthew Leifer was supported in part by MITACS and ORDCF. At Perimeter Institute, Matthew Leifer was supported in part by grant RFP1-06-006 from The Foundational Questions Institute (fqxi.org).

[1] J. Barrett, Information processing in general probabilistic theories, Phys. Rev. A. **75** 032304 (2007) (arXiv:quant-ph/0508211).
[2] H. Barnum, J. Barrett, M. Leifer and A. Wilce, Cloning and broadcasting in generic probabilistic models (2006) (arXiv:quant-ph/061129).
[3] H. Barnum, J. Barrett, M. Leifer and A. Wilce, A general no-cloning theorem, Phys. Rev. Lett. **99** 240501 (2007) (arXiv:0707.0620).
[4] H. Barnum, J. Barrett, M. Leifer and A. Wilce, Teleportation in general probabilistic theories (2008) (arXiv:0805.3553).
[5] H. Barnum, C. Fuchs, J. Renes and A. Wilce, Influence-free states on coupled quantum-mechanical systems (2005) (arXiv:quant-ph/0507108).
[6] H. Barnum, P. Gaebbler and A. Wilce, Weak self-duality and ensemble steering in general probabilistic theories, preprint.
[7] H. Barnum, E. Knill, G. Ortiz and L. Viola, Generalizations of entanglement based on coherent states and convex sets, Phys. Rev. A **68** 032308 (2003).
[8] J. Barrett and M. Leifer, The de Finetti theorem for test spaces, New J. Phys. **11** (2009), 033024 (arXiv:0712.2265).
[9] H. Barnum and A. Wilce, Information processing in convex operational theories (2009) (arXiv:0908.2352). To appear in a special issue of *Electronic Notes in Theoretical Computer Science*: Proceedings of QPL/DCM (Quantum Physics and Logic / Developments in Computational Models), Reykjavik, July 12–13, 2008.
[10] H. Barnum and A. Wilce, Ordered linear spaces and categories as frameworks for information-processing characterizations of quantum and classical theory (2009) (arXiv:0908.2354).
[11] J. F. Clauser, M. A. Horne, A. Shimony and R. A. Holt, Proposed experiment to test local hidden-variable theories, Phys. Rev. Lett. **23** 880–884 (1969).
[12] G. Chiribella, G. M. D'Ariano and P. Perinotti, Reversible realization of physical processes in probabilistic theories (2009) (arXiv:0908.1583).
[13] E. B. Davies and J. T. Lewis, An operational approach to quantum probability, Comm. Math. Phys. **17** 239-260 (1970).
[14] I. Bengtsson and M. Zyckowski, *The Geometry of Quantum States*, Cambridge University Press (2004).
[15] C. M. Edwards, The operational approach to quantum probability I, Comm. Math. Phys. **17** 207–230 (1971).
[16] D. J. Foulis and C. H. Randall, Empirical logic and tensor products, in H. Neumann, (ed.), *Interpretations and Foundations of Quantum Theory*, Bibliographisches Institut, Wissenschaftsverlag, Mannheim (1981).
[17] J. Gunson, On the algebraic structure of quantum mechanics, Communications in Mathematical Physics **6** 262–285 (1967).
[18] L. Hardy, A framework for probabilistic theories with non-fixed causal structure, J. Phys. A. **40** 3081 (2007).
[19] N. Hadjisavvas, Properties of mixtures of non-orthogonal states, Lett. Math. Phys. **5** 327–332 (1981).
[20] C. A. Hein, Entropy in Operational Statistics and Quantum Logic, Foundations of Physics **9** 751–78 (1979).
[21] L. Hughston, R. Jozsa and W. Wooters, A complete classification of quantum ensembles having a given density matrix, Phys. Lett. A **183** 14–18 (1993).
[22] K. A. Kirkpatrick, The Schrödinger-HJW theorem, Found. Phys. Lett. **19** 95–102 (2006).
[23] M. Kläy, D. J. Foulis, and C. H. Randall, Tensor products and probability weights, Int. J. Theor. Phys. **26** 199–219 (1987).
[24] M. Kläy, Einstein-Podolsky-Rosen experiments: the structure of the probability space I, Foundations of Physics **1** 205–244 (1988).
[25] G. Ludwig, *An Axiomatic Basis of Quantum Mechanics 1, 2*, Springer-Verlag (1985, 1987).
[26] G. Mackey, *Mathematical Foundations of Quantum Mechanics*, Benjamin (1963).
[27] M. Pawlowski, T. Paterek, D. Kazlikowski, V. Scarani,

A. Winter and M. Żukowski, et al. A new physical principle: Information causality (2009) (arXiv:0905.2292).
[28] S. Popescu and D. Rohrlich, Nonlocality as an axiom, Found. Phys. **24** 379 (1994).
[29] A. J. Short and J. Barrett, Strong nonlocality: A trade-off between states and measurements, arXiv:0909.2601.
[30] E. Schrödinger, Probability relations between separated systems, Proceedings of the Cambridge Philosophical Society **32** 446-452 (1936).
[31] F. W. Shultz, Journal of Combinatorial Theory A **17** 317 (1974).
[32] R. W. Spekkens, Evidence for the epistemic view of quantum states: a toy theory, Phys. Rev. A. **75** 032110 (2007).
[33] B. S. Tsirel'son, Quantum Generalizations of Bell's Inequality, Lett. Math. Phys. **4** 93 (1980).
[34] W. van Dam, Implausible consequences of superstrong nonlocality (2005) (arXiv:quant-ph/0501159).
[35] A. Wilce, Test Spaces, in D. Gabbay et al., eds., *Handbook of Quantum Logic*, North Holland (2008).
[36] A. Wilce, Formalism and interpretation in quantum theory. To appear in Foundations of Physics.
[37] A. Wilce, Tensor products of frame manuals, Int. J. Theor. Phys. **29** 805-814 (1990).
[38] A. Wilce, Topological test spaces, Int. J. Theor Phys. **44** 1227-1238 (2005).
[39] H. Ollivier and Wojciech H. Zurek, Quantum Discord: A Measure of the Quantumness of Correlations, Phys. Rev. Lett **88** 017901 (2002) (arXiv:quant-ph/0105072). V. Vedral and L. Henderson, Classical, quantum and total correlations, J. Phys. A-Math. Gen. **34** 6899–6905 (2001) (arXiv:quant-ph/0105028).
[40] Measure-theoretic classical probability theory is, in effect, the theory of systems of the form $(\mathfrak{D}, \Theta)$ where $\mathfrak{D} = \mathfrak{D}(S, \Sigma)$ is the set of all finite (respectively, countable) partitions $E = \{a_i\}$ of a measurable space S by non-empty mesurable sets $a_i \in \Sigma$, and where Θ is some closed convex set of probability measures on E. The probabilistic apparatus of states and observables associated with von Neumann algebras can be modeled in a similar way.
[41] Alternatively, this condition could be derived from some other mild conditions on test spaces, as discussed in Appendix B
[42] By [31], any compact convex set can be represented as the full state space $\Omega(\mathfrak{A})$ of some locally finite test space $\mathfrak{A}$.
[43] To be a bit more precise: a quantum state is the quadratic form associated with a density operator. We shall routinely identify a density operator ρ with its quadratic form, writing $\rho(x)$ for $\langle \rho x, x\rangle$ where x is a unit vector on $\mathbf{H}$.
[44] A proof that the von Neumann entropy minimizes mixing entropy can be found in [14]. The key observation is that the mixing coefficients for any ensemble for ρ can be obtained from the eigenvalues of ρ by a doubly stochastic matrix (Shrodinger's Lemma), which can only increase entropy. An easier version of the same argument (also noted by Hein [20]) shows that the spectral decomposition also minimizes measurement entropy.
[45] This example, well-known in the quantum-logical literature, has a fairly concrete interpretation in terms of a firefly in a three-chambered triangular box. See [35] for details.
[46] Quantum theory does just that: the rule is that the joint measurements and states correspond respectively to maximal sets of pairwise orthogonal projections and density operators on the tensor product of the individual Hilbert spaces.
[47] Clearly, the content of the DPI depends on the definition of mutual information, which in turn depends on the definition of entropy used. As before, we continue to define mutual information in terms of the measurement entropy, as in Definition 6. But note that if the DPI fails for one definition of mutual information, the corresponding condition may hold for another - indeed this may be a reason to prefer the latter. In this paper, we omit a formal treatment of "processes".
[48] This does not hold in all theories.
[49] Recall that the joint test space for a classical and general system is always assumed to be the Foulis-Randall product, and that this includes measurements of the form: measure F first and, depending on the outcome, measure B.

APPENDIX A: NON-CONCAVITY OF MIXING ENTROPY

In this appendix we prove Theorem 2, which states that the mixing entropy is not concave for nonsimplicial polytopes. As a preliminary to the proof, we state some basic definitions and facts that we will use. A *face* of a convex set C, which is a set $F \subseteq C$ such that every $x \in C$ that can appear in a convex decomposition of something in F, is also in F. A *maximal face* of C is one that is not a proper subset of any face in C other than C itself. An *exposed face* of C is a subset of C that is the intersection of C with a hyperplane supporting it (such a subset is easily shown to be a face). All faces of a polytope are exposed, and the maximal ones have affine codimension 1, i.e. their spans are affine hyperplanes. We denote the affine space generated by a set S by $\mathrm{aff}(S)$, the linear span of S by $\mathrm{lin}(S)$, and the cone generated by S (i.e. the set of nonnegative linear combinations of elements of S) by $\mathrm{cone}(S)$. Note that when a subset of a real vector space contains 0, $\mathrm{aff}(S) = \mathrm{lin}(S)$. The *relative interior* of a convex compact set C is the interior of C when it is considered as a subset of $\mathrm{aff}(C)$. Finally, we'll use the term *Z-ball*, where Z is an affine subspace of the ambient vector space, to mean a subset of Z that is a ball in Z.

The proof relies on the following lemma, proven below.

Lemma 5 *The mixing entropy fails to be concave for any d-dimensional nonsimplicial polytope all the maximal faces of which are (d-1)-dimensional simplices.*

We begin by proving the theorem.

Proof. (Theorem). First note that any counterexample

to concavity of the mixing entropy in a polytope S will also be a counterexample in a polytope S' that has S as a face. This follows from the fact if S is a face, only states in S can appear in convex decompositions of states in S. The proof of the theorem is by induction.

Suppose as our induction hypothesis that the mixing entropy fails to be concave for nonsimplicial polytopes in dimension d. For every polytope in dimension $d+1$ either (i) every maximal face is simplicial, or (ii) there is a maximal face that is nonsimplicial. If case (ii) applies, then there is a face that constitutes a nonsimplicial polytope of dimension d and by our induction hypothesis, the mixing entropy fails to be concave for this face. If case (i) applies, then the polytope satisfies the conditions of the lemma and the mixing entropy fails to be concave by virtue of the lemma.

To complete the inductive argument we need to show that the mixing entropy fails to be concave for nonsimplicial polytopes in dimension $d = 2$, the lowest dimension in which there exist nonsimplicial polytopes. This follows from the fact that all of the maximal faces of a 2-dimensional nonsimplicial polytope are line segments, which are simplices, so that the conditions of the lemma apply. □

We now prove the lemma.

Proof. (Lemma). Suppose S is a d-dimensional polytope that satisfies the conditions of the lemma, that is, it is nonsimplicial, but all of its maximal faces are simplicial. In this case, one can always find two maximal faces ($(d-1)$-dimensional simplices), F_1 and F_2, whose intersection, $F_1 \cap F_2$, is a $(d-2)$-dimensional simplex. We define V_1 to be the vertex of F_1 that is not contained in $F_2 \cap F_2$. V_2 is defined similarly. Let ρ_1 be the barycenter of F_1, ρ_2 the barycenter of F_2 and ρ_3 the barycenter of $F_1 \cap F_2$. The figures provide examples of pairs of such faces in different dimensions.

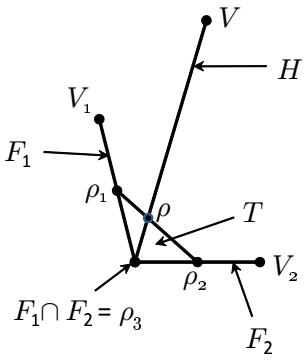

FIG. 7: Example of failure of concavity for a 2d nonsimplicial polytope.

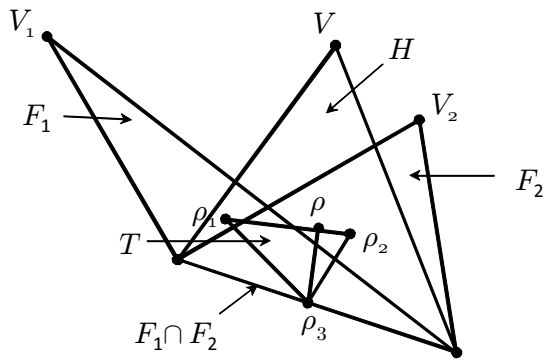

FIG. 8: Example of failure of concavity for a 3d nonsimplicial polytope. Here ρ is in the interior of H.

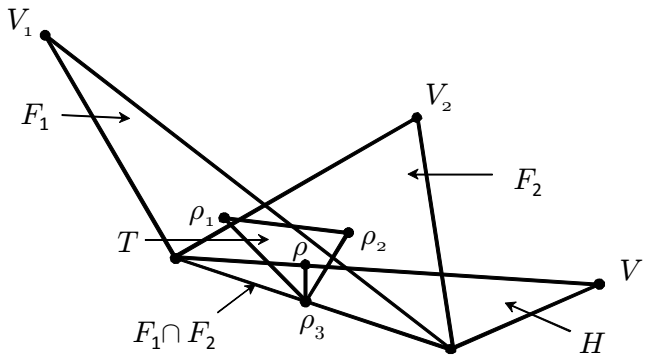

FIG. 9: Example of failure of concavity for a 3d nonsimplicial polytope. Here ρ is on the boundary of H.

Let V be a vertex of S that is not contained in F_1 or in F_2. Such a vertex always exists because if it did not, then the total number of vertices in S would be $d+1$ and S would be a simplex, contrary to hypothesis.

Define the $(d-1)$-dimensional polytope H to be the convex hull of $F_1 \cap F_2$ and V. Note that H is a simplex.

Define T to be the triangle with vertices ρ_1, ρ_2 and ρ_3.

Define L to be the intersection of T and the $(d-1)$-dimensional polytope H. L is a line segment; we defer establishing this to the end of the proof, since it is somewhat technical.

Finally, we define the state ρ for which the mixing entropy will fail to be concave. It is defined as the second vertex of L, that is, L is the line segment extending from ρ_3 to ρ.

Now the proof proceeds differently depending on whether ρ is in the interior or in the boundary (relative to $\mathrm{aff}(H)$) of H.

i) ρ is in the relative boundary of H.

In this case, ρ lies on a face of H. Because H is a simplex of dimension $d-1$, every such face has $d-1$ vertices and consequently the mixing entropy of ρ satisfies

$$S(\rho) \leq \log(d-1). \tag{A1}$$

By definition, $\rho \in T$, so that it is a convex combination of ρ_1, ρ_2 and ρ_3,

$$\rho = p_1\rho_1 + p_2\rho_2 + p_3\rho_3, \tag{A2}$$

where the p_i form a probability distribution. Because ρ_1 (ρ_2) is the barycenter of F_1 (F_2), which has d vertices, its mixing entropy is

$$S(\rho_1) = S(\rho_2) = \log d. \tag{A3}$$

while because ρ_3 is the barycenter of $F_1 \cap F_2$ with $d-1$ vertices, we have

$$S(\rho_3) = \log(d-1). \tag{A4}$$

Recalling that L is 1-dimensional, we know that $\rho \neq \rho_3$, or equivalently, $p_1 + p_2 > 0$, which implies that

$$\sum_{i=1}^{3} p_i S(\rho_i) > \log(d-1). \tag{A5}$$

From Eqs. (A1) and (A5) we infer that

$$S(\rho) < \sum_i p_i S(\rho_i), \tag{A6}$$

that is, the mixing entropy fails to be concave.

ii) ρ is among the relative interior points of H.

In this case,

$$\rho = p_1\rho_1 + p_2\rho_2, \tag{A7}$$

that is, ρ lies on the line segment defined by ρ_1 and ρ_2. The proof of (A7) is by contradiction. Suppose that ρ lies in the relative interior of T as well as in the relative interior of H. Then, there is an $\mathrm{aff}(T)$-ball B_1 around ρ contained in the relative interior of T and an $\mathrm{aff}(H)$-ball around ρ contained in the interior of H. $B_1 \cap B_2$ is a line segment $L_\rho \subset L$ with midpoint ρ. But the fact that ρ is the midpoint of L_ρ contradicts the fact that it is an extremal point of L.

It follows from Eq. (A7) and the fact that ρ_1, ρ_2 are barycenters of $d-1$-dimensional simplices that

$$\sum_i p_i S(\rho_i) \,(= p_1 S(\rho_1) + p_2 S(\rho_2)) = \log d. \tag{A8}$$

Next, we show that ρ cannot be the barycenter of H. We begin by demonstrating that ρ is the barycenter of $H' \equiv \mathrm{conv}(F_1 \cap F_2, V')$, where $V' \equiv p_1V_1 + p_2V_2$ and where $\mathrm{conv}(S, S')$ is the convex hull of $S \cap S'$. Letting $\mathrm{bary}(F)$ denote the barycenter of F, the proof is as follows, writing $x_i, i \in \{1, ..., d-1\}$, for the $d-1$ vertices of $F_1 \cap F_2$.

$$\begin{aligned}
\rho &= p_1 \mathrm{bary}(F_1) + p_2 \mathrm{bary}(F_2) && \text{(A9)}\\
&= p_1 \frac{1}{d}\Big(\sum_{i=1}^{d-1} x_i + V_1\Big) + p_2 \frac{1}{d}\Big(\sum_{i=1}^{d-1} x_i + V_2\Big) && \text{(A10)}\\
&= \frac{1}{d}\Big(\sum_{i=1}^{d-1} x_i + p_1V_1 + p_2V_2\Big) && \text{(A11)}\\
&= \mathrm{bary}(\mathrm{conv}(F_1 \cap F_2, p_1V_1 + p_2V_2)) && \text{(A12)}\\
&= \mathrm{bary}(H'). && \text{(A13)}
\end{aligned}$$

However, $H' = \mathrm{conv}(F_1 \cap F_2, V')$ has a different barycenter from $H \equiv \mathrm{conv}(F_1 \cap F_2, V)$ because V is distinct from V'. The latter follows from the fact that V, V_1 and V_2 are all vertices of the nonsimplicial polytope S, and consequently V cannot be in $\mathrm{conv}(F_1 \cap F_2, V_1, V_2)$, unlike V' which is.

Given that $\rho \in H$ but is not at its barycenter, and given that H has d vertices, we have

$$S(\rho) < \log d. \tag{A14}$$

From Eqs. (A8) and (A14) we infer the failure of concavity of the mixing entropy.

We finish the proof by establishing the claim that L, defined above, is a line segment. First note that because it is an intersection of convex compact sets, it is compact and convex. Because $\dim(\mathrm{aff}(T)) = 2$ and $\dim(\mathrm{aff}(H)) = d-1$, $\mathrm{aff}(T) \cap \mathrm{aff}(H)$ is one or two dimensional. For it to be two dimensional would require $T \subset \mathrm{aff}(H)$, implying $\rho_1, \rho_2 \in \mathrm{aff}(H)$, and hence since $F_1 \cap F_2 \subset H$, that V_1, V_2 lie in the hyperplane $\mathrm{aff}(H)$. That contradicts the assumption that F_1, F_2 are distinct maximal faces.

Since $L \subset \mathrm{aff}(T) \cap \mathrm{aff}(H)$, L is at most one-dimensional. To show it is *at least* 1-dimensional we begin by observing that because they are subsets of S, both T and H lie in the "tangent wedge" W to S at ρ_3, i.e. the intersection of the halfspaces $\mathrm{aff}(F_1)_+$ and $\mathrm{aff}(F_2)_+$. Here $\mathrm{aff}(F_i)_+$ is defined to be the closed half-space to the polytope S's side of $\mathrm{aff}(F_i)$. In fact, V lies in the *interior* of W because if it lay in $\mathrm{aff}(F_1)$ or $\mathrm{aff}(F_2)$, our assumption that all maximal faces were simplices would be violated. Viewing ρ_3 as the origin of a real linear space, and noting that $\mathrm{lin}T$ and $\mathrm{lin}(F_1 \cap F_2)$ are complementary subspaces (they span the space and intersect only at 0, i.e. ρ_3) we can decompose V in a unique way into a component in $\mathrm{lin}(T)$ and a component in the edge, $\mathrm{lin}(F_1 \cap F_2)$, of the tangent wedge.

Let q be the linear projection with kernel $\mathrm{lin}(F_1 \cap F_2)$ and image $\mathrm{lin}(T)$. For any set X such that $X = X +$

$\mathrm{lin}(F_1 \cap F_2)$, $q(X) = \mathrm{lin}(T) \cap X$. Both W and $\mathrm{aff}(H)$ satisfy this condition. As already noted, $V \in \mathrm{int}\, W$; this is equivalent to $q(V)$ being in the relative interior of $\mathrm{cone}(T)$. Since $V \in H$, $q(V)$ is also in $\mathrm{aff}(H)$. Therefore $\mathrm{cone}(T) \cap \mathrm{aff}(H)$ is an interior ray r of $\mathrm{cone}(T)$. Now, since ρ_3 is the barycenter of $F_1 \cap F_2$, there is a $d-2$-dimensional $\mathrm{aff}(F_1 \cap F_2)$-ball B_0 around ρ_3 contained entirely in $F_1 \cap F_2$. Furthermore, B_0 is the intersection of a $d-1$-dimensional $\mathrm{aff}(H)$-ball B around ρ_3, with $F_1 \cap F_2$. $\mathrm{aff}(F_1 \cap F_2)$ divides $\mathrm{aff}(H)$ into halfspaces, and the H-side half-ball $B \backslash B_0$ is contained entirely in the relative interior of H. Since, as established near the beginning of our argument, $\mathrm{aff}(T) \cap \mathrm{aff}(H)$ is a line in $\mathrm{aff}(H)$, and we now know that while it contains ρ_3 it is not entirely contained in $F_1 \cap F_2$, it must intersect $B \backslash B_0$. Its intersection with $B \backslash B_0$ is contained in H. By choosing B small enough, we can ensure that this intersection is also contained in T. This is obvious from two-dimensional geometry. To be slightly more explicit, the facts that r, i.e. the half of $\mathrm{aff}(H) \cap \mathrm{aff}(T)$ on the H-side of $\mathrm{aff}(F_1 \cap F_2)$, is interior to $\mathrm{cone}(T)$, $\mathrm{cone}(T)$ is generated by T, and T is closed under multiplication by scalars in $[0, 1]$, ensure this. Since $\mathrm{aff}(T) \cap \mathrm{aff}(H) \cap B$ is contained in both H and T, it is contained in L; since it is one-dimensional, so is L and so, since L is a compact convex set, L is a line segment. □

In generalized theories we can define (cf. also [7], where analogous quantities for convex-sets-based theories were defined, and their failure to be concave in general was also observed) measurement-entropy-like quantities H_T based on *any* function T that (like entropy) is Schur-concave and defined on finite lists of classical probabilities. For a state ρ, $H_T(\rho)$ is defined as the infimum over tests of the value of T on the probabilities for the results of the test. We define U_d for positive integers d as the uniform distribution on d alternatives. The same proof as before (with $F(U_d)$ in place of $\log d$) gives us:

Proposition 1 *For any T whose value on U_{d+1} is strictly greater than its value on U_d, for all d, for example any strictly Schur-concave T, the only polytopes on which H_T is concave are simplices.*

APPENDIX B: ENTROPY AND QUANTUM AXIOMATICS

That mixing and measurement entropies coincide, as they do in classical and quantum theory, has powerful consequences for the structure of a probablistic model and, perhaps even more profoundly, for the structure of a probabilistic *theory*. As already noted, it implies that mixing entropy is concave, which places sharp restrictions on the geometry of state spaces. It also figures importantly in our derivation, in Section V, of information causality. In this Appendix, we explore some further consequences of monoentropicity, and also suggest some other postulates, the physical content of which may be clearer, that enforce this property.

It will be helpful to impose some mild restrictions on the models we consider. (These are satisfied by all of the examples discussed earlier.) First, we want to have enough analytic structure to guarantee that measurement entropies will be well-behaved. Accordingly, in this Appendix we shall require of all models $A = (\mathfrak{A}, \Omega)$ that Ω be a compact, finite-dimensional convex set, as already assumed in Section II; additionally, we make the following technical, but reasonable and fairly weak, assumptions:

(i) The total outcome-set X is compact in some Hausdorff topology that makes every state $\alpha \in \Omega^A$ continuous as a function $\alpha : X \to [0, 1]$.

(ii) Write $x \perp y$ to mean that outcomes x and y are distinct and jointly testable, i.e., there exists a test $E \in \mathfrak{A}$ containing them both. We require that $\perp$ be closed as a subset of $X \times X$.

(iii) $\mathfrak{A}$ is compact in the standard topology it inherits from X (as explained below).

Conditions (i) and (ii) have a certain *a priori* plausibility, and, indeed, are often satisfied in practice: see [35] for examples of large classes of test spaces satisfying them. Condition (iii) requires some further justification. Conditions (i) and (ii) make $\mathfrak{A}$ a *topological test space* [35, 38]. With X compact, as in condition (i), $\mathfrak{A}$ has finite rank [35], Lemma 204. This allows us to topologize the set $\mathfrak{A}$ of tests as a quotient of a suitable subspace of X^n, where n is the rank of $\mathfrak{A}$. We call this the *standard topology* on $\mathfrak{A}$. One can show ([35], Prop. 211) that $\mathfrak{A}$ can be enlarged so as to become compact in this topology, without change to its rank or to its space of continuous states. So condition (iii) is relatively harmless. In fact, if $\mathfrak{A}$ is *uniform*, meaning that all tests have the same number of outcomes, condition (iii) is automatically satisfied, given (i) and (ii).

It is not difficult to show that $H_E(\alpha)$ is continuous as a function of $E \in \mathfrak{A}$ (see [35], Lemma 210). Hence, for every state $\alpha \in \Omega$, there exists a test E for which $H_E(\alpha) = H(\alpha)$. This justifies the assumption to this effect made in Section III.

We can now characterize those states having zero measurement or mixing entropy.

Lemma 6 *Let α be a state of a system $A = (\mathfrak{A}, \Omega)$ satisfying the standing assumptions just discussed.*

(a) $H(\alpha) = 0$ *iff there exists an outcome* $x \in X$ *with* $\alpha(x) = 1$.

(b) If $S(\alpha) = 0$, *then* α *is the limit of a sequence of pure states of* Ω.

Proof. (a) "if" follows immediately from the definition, with E any test containing x; "only if" from the fact (established just above the statement of the Lemma) that $H(\alpha) = H_E(\alpha)$ for some $E \in \mathfrak{A}$. For (b), note that if $\vec{p} = (p_1, ..., p_n)$ is a discrete probability distribution with $H(\vec{p}) < \epsilon$, then $\max\{p_i\} > 2^{-\epsilon}$. Now if $S(\alpha) = 0$, we can find, for any sequence ϵ_k decreasing to 0, a sequence of pure-state ensembles $\{p_{i,k}\alpha_{i,k} | i = 1, ..., n_k\}$ for α (so that $\alpha = \sum_{i=1}^{n_k} p_{i,k}\epsilon_{i,k}$ for every k) with $H(\vec{p_k}) < \epsilon_k$. Ordering each ensemble so that $p_{1,k} = \max\{p_{i,k} | i = 1, ..., n_k\}$, we find, as above, that $p_{1,k} > 2^{-\epsilon_k}$. Since $\alpha = p_{1,k}\alpha_{1,k} + \sum_{i=2}^{n} p_{i,k}\epsilon_{i,k}$, we have $\alpha > p_{1,k}\alpha_{1,k}$ in the pointwise order on $X = \bigcup \mathfrak{A}$; consequently, $\|\alpha - p_{1,k}\alpha_{1,k}\| = (\alpha - p_{1,k}\alpha_{1,k})(u) = 1 - p_{1,k}$. Thus, $\alpha = \lim_k \alpha_{1,k}$. □

The converse to part (b) would trivially be true if the mixing entropy were continuous on the convex set Ω. However, as Example 2 of the main text shows, it need not be.

Call a model $A = (\mathfrak{A}, \Omega)$ *unital* iff for every outcome $x \in X := \bigcup \mathfrak{A}$, there exists at least one state α with $\alpha(x) = 1$. If this state is unique—and therefore pure—for every x, we say that A is *sharp*. In this case, we write ϵ_x for the unique state with $\epsilon_x(x) = 1$. In the literature of quantum axiomatics, sharpness has sometimes been taken as an axiom (sometimes called *Gunson's Axiom*) [17]. Lemma 6 has the following corollary:

Corollary 3 *Suppose* A *is monoentropic, and that the set of pure states in* A *is closed. Then* A *is sharp.*

Proof. If α is a pure state, then $H(\alpha) = S(\alpha) = 0$. By Lemma 6, there exists a measurement outcome x with $\alpha(x) = 1$. On the other hand, if $\alpha(x) = 1$, then $S(\alpha) = H(\alpha) = 0$, whence, again by Lemma 6, α is the limit of a sequence of pure states, say $\epsilon_n \to \alpha$. By assumption, the set of pure states is closed, so α is pure. Since the set of states assigning unit probability to x is convex, it follows that α is the unique such state. □

While the condition that the set of pure states be closed is not totally innocent (consider, e.g., Example 2 above), neither is it unreasonable. For example, it will be satisfied if there exists a compact group of symmetries of the state space that acts transitively on the pure states.

The condition that measurement and mixing entropies coincide also places some constraints on how systems compose:

Lemma 7 *Suppose that* $AB = (\mathfrak{C}, \Omega^{AB})$ *is a composite (in the sense of Section II) of systems* $A = (\mathfrak{A}, \Omega^A)$ *and* $B = (\mathfrak{B}, \Omega^B)$. *Suppose that* A, B *an* AB *have closed sets of pure states, and are monoentropic. If* Ω^{AB} *contains an entangled pure state, then* $\mathfrak{C}$ *must contain a non-product outcome.*

Proof. By the previous Lemma, A, B and AB are sharp. If $x \in \bigcup \mathfrak{A}$ and $y \in \bigcup \mathfrak{B}$ are outcomes of A and B, respectively, and ϵ_x, ϵ_y and ϵ_{xy} are the unique pure states making x, y and xy certain, then $\epsilon_{xy} = \epsilon_x \otimes \epsilon_y$. Now if ρ is a pure entangled state in Ω^{AB}, then $S(\rho) = 0$. If $H = S$, then $H(\rho) = 0$, whence, $\rho = \epsilon_z$ for some outcome $z \in \bigcup \mathfrak{C}$. If z is a product outcome, say $z = xy$, then $\rho = \epsilon_x \otimes \epsilon_y$ – a contradiction. □

We now consider whether the condition that $H = S$ can be derived from more physically transparent considerations.

Definition 8 *A probabilistic theory has the* pure conditioning property *iff, for every pair of systems* $A = (\mathfrak{A}, \Omega^A)$ *and* $B = (\mathfrak{B}, \Omega^B)$, *every pure state* ω *of* AB, *and all outcomes* x *of* A *and* y *of* B, *the conditional states* $\omega^{B|x}$ *and* $\omega^{A|y}$ *are pure.*

Lemma 8 *If a theory satisfies the pure conditioning property, then for any pure bipartite state* ω *on a composite* AB, *we have* $S(\omega^B) \leq H(\omega^A)$ *and* $S(\omega^A) \leq H(\omega^B)$.

Proof. Let ω be a pure bipartite state. Pick an observable E minimizing measurement entropy for ω_1, so that $H(\omega_1)$ is the Shannon entropy $H_E(\omega^A) := -\sum_{x \in E} \omega^A(x) \log(\omega^A(x))$. We have $\omega^B = \sum_{x \in E} \omega^A(x)\omega^{B|x}$. By PC (and the assumption that ω is pure), the conditional states $\omega^{B|x}$ are pure. By definition, $S(\omega^B)$ is the minimum Shannon entropy of the mixing coefficients in any pure-state ensemble for ω^B, so $S(\omega^B) \leq H_E(\omega^A) = H(\omega^A)$. By the same argument, $S(\omega^A) \leq H(\omega^B)$. □

Definition 9 *A theory has the* steering property *iff, for every pair of systems* A *and* B, *every pure bipartite state* ω *of* AB steers *its marginals, in the sense that for any convex decomposition* $\omega^B = \sum_i p_i\beta_i$, *with* β_i *pure and distinct from each other, there is a test* $E = \{a_i\}$ *of* A *with* $\beta_i = \omega^{B|a_i}$, *and similarly for* ω^A.

The term "steering" is due to Schrödinger [30], who showed that quantum theory is steering; further proofs and extensions are in Hadjisavvas [19] and Hughston, Jozsa, and Wootters [21]; a survey is [22].

Lemma 9 *If a theory has the steering property, then for every pure bipartite state* ω, $H(\omega^A) \leq S(\omega^B)$.

Proof. For any $\epsilon > 0$, choose a convex decomposition $\omega^B = \sum_i p_i \beta_i$ of ω^B into pure states β_i, with $S(\omega^B) > H(p_i) - \epsilon$. Since the state ω is steering, there exists a test $E = \{x_i\}$ with $\omega^{B|x_i} = \beta_i$, whence $p_i = \omega^A(x_i)$. It follows that $S(\omega^B) > -\sum_i p_i \log(p_i) - \epsilon = H_E(\omega^A) - \epsilon$. Since ϵ is arbitrary, $S(\omega^B) \geq H(\omega^A)$. □

Definition 10 *A pure state α in an abstract probabilistic theory is* purifiable *iff for every state α on a system A, there exists a pure bipartite state ω – a* purification *of α – on a composite AB, with B a copy of A, with $\omega^A = \omega^B = \alpha$. An abstract probabilistic theory has the* purifiability property *iff every state in the theory is purifiable.*

Quantum mechanics has the purifiability property. D'Ariano *et al.* [12] have considered a condition very similar to purifiability as a potential axiom for quantum theory, and have shown that many other features of quantum theory follow from it. From the Lemmas above, we have

Proposition 2 *A theory that has the pure conditioning, steering and purifiability properties is monoentropic.*

APPENDIX C: LINEARIZED TEST SPACE MODELS, ORDERED LINEAR SPACE MODELS, AND ENTROPY

The apparatus of states on test spaces can be linearized, as follows. If $A = (\mathfrak{A}, \Omega)$, with total outcome space $X = \bigcup \mathfrak{A}$, let $V(A)$ denote the span of Ω in $\mathbb{R}^X$, regarded as an ordered real vector space with positive cone $V_+(A) = \{\alpha \in V(A) | \alpha(x) \geq 0 \forall x \in X\}$. Every outcome $x \in X$ defines a positive linear evaluation functional $f_x \in V^*(\Omega)$ by $f_x(\mu) = \mu(x)$ for all $\mu \in V(A)$. Moreover, one has $\sum_{x \in E} f_x = u$, where u is the unique functional taking the constant value 1 on Ω. Abstracting, one defines an *effect* to be a positive linear functional $f \in V^*(A)$ with $0 \leq f(\alpha) \leq 1$ for all $\alpha \in \Omega$ (equivalently, $0 \leq f \leq u$); an *observable* on A is a sequence $f_1, ..., f_n$ of effects with $\sum_i f_i = u$.

From this point of view, the structure of the test space is essentially a privileged set of observables – an additional structure that (like a preferred basis for a vector space) may or may carry some useful information, or may simply be a computational convenience. For example, if $A(\mathbf{H}) = (\mathfrak{F}(\mathbf{H}), \Omega(\mathbf{H}))$ is a quantum system, $V(A)$ is the space of quadratic forms associated with – but one might as well say, the space *of* – Hermitian operators on $\mathbf{H}$, and V^* is essentially the same space, under the duality $a(\rho) = \mathrm{Tr}(\rho a)$. In particular, an effect is a positive operator between 0 and $\mathbf{1}$, and an observable is essentially a discrete POVM. The convex sets, or ordered linear spaces, formalism takes this kind of combination of a convex state space and a set of effects in the dual cone to the state space, as primary. Roughly, a convex model is defined by taking a convex compact set of states as a base for a cone $V(\Omega)_+$ of unnormalized states, and a cone of "unnormalized allowed effects" that is a closed subcone $V_+^\sharp$, containing u in its interior, of the dual cone $V^*(\Omega)_+$ of all effects. u is defined by the condition $u(\Omega) = 1$, and the interval $[0, u]$ according to the ordering defined by $V^\sharp$ is the set of effects allowed in the theory. When $V^\sharp = V(\Omega)_+$, the model is called *maximal* (or sometimes *saturated* [10]). If the model is constructed from a test space, one will usually want to choose $V^\sharp$ to contain the effects associated with all outcomes in the test space.

Two natural distinguished classes of effects are the *ray-extremal* ones, that is effects that lie on extremal rays of the cone generated by effects, and *atomic* effects, i.e., maximal effects in extremal rays (equivalently, ray-extremal effects that are also extremal in the convex set $[0, u]$ of effects). We may define the measurement entropy as the infimum of entropies obtainable by measuring observables consisting of ray-extremal elements, or alternatively as the infimum of entropies obtainable by measuring observables consisting of atomic effects. Intuitively, the observables consisting of *ray-extremal* effects are maximally fine-grained. Ray-extremal effects cannot be further refined by decomposing them as sums of other effects. Although they can be decomposed as sums of shrunken versions of themselves, intuitively this cannot provide any additional information about the system being measured. Certainly in the case of atomic effects, and probably with some care and relabeling in the case of ray-extremal effects (which unlike atomic effects may appear more than once in a given observable), the measurements with such outcomes can be organized into distinguished test spaces associated with a given convex-sets model, so the test space framework we use in the main text will probably cover this natural possibility, although the additional assumptions we make to obtain particular results will need to be checked for these cases. In the case of ray-extremal effects, the infimum in the definition of measurement entropy will likely not be changed if we omit measurements in which an effect appears more than once; the measurements without repetitions should be easier to organize into a test space. Linearization and the application of one of these definitions may well remove pathologies in measurement entropy that are associated with some test space/state space models. The spirit of the definition of measurement entropy via an infimum suggests excluding tests that are not maximally fine grained when viewed from the convex states perspective, as the above definitions do. Passing to these definitions may also remove pathologies that might arise when the set of distinguished observables associated with tests has an irregular relationship to a state space whose underlying geometry is quite regular.